DURGARAJ CHAUHAN

AI Engineer | Full Stack Developer

Mumbai, Maharashtra 400030 | +91-8268874907 | durgarajchauhan@gmail.com

LinkedIn: linkedin.com/in/durgaraj-chauhan | GitHub: github.com/DurgarajC07 | Portfolio: durgarajchauhan.vercel.app

PROFESSIONAL SUMMARY

AI/ML Engineer with 2.5+ years of experience building production-grade Voice AI, Vision AI, and GenAI systems. Expertise in LLMs (GPT-4, Gemini, Mistral), computer vision (YOLO, face recognition), and real-time inference pipelines. Skilled in Python/FastAPI microservices, RAG architectures, fine-tuning transformers, and deploying scalable ML solutions on AWS/Azure. Strong full-stack capabilities with React.js and Node.js for end-to-end AI product development.

TECHNICAL SKILLS

AI/ML: LLMs (GPT-4o, Gemini, Mistral, Claude) | RAG | Fine-tuning | Prompt Engineering | Agentic AI | Langchain | Vector Databases | Transformers | Embeddings | Semantic Search

Computer Vision: YOLO | OpenCV | Face Recognition | Object Detection | Video Analytics | Real-time Inference | RTSP Streaming

Speech AI: STT (Sarvam) | TTS (ElevenLabs, Sarvam TTS) | Audio Processing

Backend: Python | FastAPI | Django | Flask | Node.js | Nest.js | PHP Laravel | RESTful APIs | WebSockets | Microservices

Frontend: React.js | JavaScript | HTML5 | CSS3 | Bootstrap | Tailwind CSS | Shadcn UI

Databases: PostgreSQL | MySQL | SQLite | Redis | ChromaDB | Qdrant | Vector DBs

Cloud & DevOps: AWS (S3, EC2, Lambda) | Azure (Video Indexer, Cognitive Services) | Docker | CI/CD | Git

Tools: Postman | Bitbucket | GitLab | Jupyter | Pandas | NumPy | Scikit-learn | Copiliot

PROFESSIONAL EXPERIENCE

SR. SOFTWARE ENGINEER | Anvex AI Technologies Pvt Ltd, Mumbai January 2025 - Present

- Architected real-time Vision AI pipeline using YOLO and OpenCV with socket-based alert delivery, processing 30+ FPS from live RTSP streams with <500ms inference latency for industrial safety monitoring
- Built production Voice AI system handling 500+ concurrent telephony calls using FastAPI, WebSockets, and async processing, achieving 95% conversation completion rate with multi-turn dialogue management
- Integrated LLM orchestration layer (GPT-4o, Mistral) with RAG architecture using vector databases (ChromaDB) for context-aware responses, reducing hallucination by 40% in domain-specific conversations
- Developed multilingual STT/TTS pipeline supporting 10+ languages (Sarvam + ElevenLabs), optimizing audio chunking and streaming to achieve <1.5s end-to-end voice response latency
- Created AI-powered document intelligence platform processing PDFs, DOCX, and images using Gemini Vision API, extracting structured data from 1000+ forms with 92% accuracy, integrated with batch voice automation
- Implemented state management and observability for call lifecycle tracking, reducing system failures by 35% through comprehensive logging, metrics collection, and automated failure recovery mechanisms

- Optimized inference pipelines with batching, caching, and async scheduling, improving throughput by 3x while reducing cloud costs by 25% through efficient resource utilization

JR. WEB DEVELOPER | Reboot Technology Pvt Ltd, Mumbai June 2023 - January 2025

- Built computer vision services for face recognition and real-time object detection using YOLO, processing 20+ concurrent RTSP streams with frame skipping and multi-threaded inference optimization
- Integrated Azure Video Indexer for automated video content analysis, extracting faces, objects, OCR, and scene metadata through REST APIs, enabling searchable video intelligence for 10K+ hours of content
- Developed QR-based attendance system with facial verification, screen recording Chrome extension with cloud sync, and secure OAuth2.0 authentication plugins handling 5000+ daily active users
- Integrated payment gateways (Stripe) and communication APIs (Twilio 2FA, SMS), DocuSign for e-signatures, and AWS S3 for scalable media storage
- Designed RESTful backend APIs with Laravel/FastAPI serving 50K+ daily requests, implemented Redis caching reducing DB load by 60%, and built subscription management with automated billing workflows
- Deployed Dockerized microservices on AWS EC2 with load balancing, achieving 99.5% uptime and <200ms average API response time

EDUCATION

Bachelor of Engineering | Saraswati College of Engineering, Kharghar 2020 - 2023

Percentage: 78.34%

Diploma in Computer Engineering | Saraswati Institute of Technology 2017 - 2020

Percentage: 90.91%

CERTIFICATIONS & ACHIEVEMENTS

- Python Programming | Docker Containerization | MVC Architecture | Laravel Master | PSD to HTML
- Built 10+ production AI/ML projects with 100K+ combined user interactions
- Active contributor to AI/ML communities and open-source projects